Perceptual and Motor Skills, 1965, 20, 311-316. © Southern Universities Press 1965

CONFIDENCE RATINGS AND LEVEL OF PERFORMANCE ON A JUDGMENTAL TASK[1]

RAYMOND S. NICKERSON AND CHARLES C. MCGOLDRICK, JR.

Decision Sciences Laboratory Electronic Systems Division
Bedford, Massachusetts

Summary.—A 4-alternative forced-choice test was administered to 96 *S*s. *S*'s task was to attempt to select the correct alternative from each test item and to indicate his degree of confidence in his choice on a 5-point rating scale. The objective was to compare the confidence assignments of *S*s who did relatively well on the primary judgmental task with those of *S*s who did poorly. It was found that *S*s who performed poorly on the primary task (LP *S*s) tended on the average to use lower confidence ratings than *S*s who did relatively well (HP *S*s). Although few used either high or low ratings exclusively, all *S*s tended to use one end of the confidence scale much more frequently than the other. However, whereas HP *S*s were fairly consistent in using the high end of the scale, LP *S*s were about evenly divided between those using the high end and those using the low. For both groups, performance tended to be monotonically related to expressed confidence. In terms of measures developed by Adams and Adams, HP *S*s made more "realistic" confidence judgments than did LP *S*s; however, there was no striking difference between groups in terms of differences in performance associated with step increases in expressed confidence.

A positive monotonic relationship between expressed confidence and objective performance measures has been obtained with a variety of judgmental tasks (e.g., Henman, 1911; Pollack & Decker, 1958; Carterette & Cole, 1959; Nickerson & McGoldrick, 1963). Unfortunately, the same confidence-performance relationship could be obtained from pooled data if either (a) each *S* used each confidence rating equally often, assigning high ratings relatively more frequently to correct judgments, or (b) the most frequently correct *S*s used only high ratings while *S*s less frequently correct used only low ones. A cursory inspection of the data of previous experiments makes it obvious that neither of these response patterns is generally the case. Typically, *S*s do not use each confidence rating with equal frequency, nor do they restrict themselves exclusively to one or the other end of the scale. What is not obvious is whether *S*s who perform relatively well on the primary judgmental task distribute their confidence ratings differently, e.g., more realistically, than do those whose primary task performance is low.

METHOD

Stimulus material and procedural details have been described in a previous report (Nickerson & McGoldrick, 1963). Only set M of these materials was

[1] This report is identified as Electronics Systems Division Technical Documentary Report No. ESD-TDR-64-236. Further reproduction is authorized to satisfy the needs of the U. S. Government. The authors wish to express their appreciation to Alan Meskil who prepared the figures.

used in the present experiment. Briefly, *S*s were given a 100-item 4-alternative forced-choice test, each item consisting of the names of 4 states. *S*'s task was to identify the largest state (area) in each item and to indicate his degree of confidence in his choice on a 5-point scale ranging from "certain I'm correct" to "pure guess." Usual randomization and counterbalancing procedures were followed to minimize effects of patterning or position biases. Seventy-two paid undergraduate college students served as *S*s. Their data were pooled with those of the 24 Group M *S*s of the previous experiment to yield an *N* of 96.

Results and Discussion

*S*s were rank-ordered in terms of their performance on the primary judgmental task. The 24 most frequently, and the 24 least frequently, correct *S*s were designated as high performance (HP) and low performance (LP) groups, respectively. The mean percent correct size judgments for the former group was 64 and for the latter, 33.

A mean confidence rating was calculated for each *S* by multiplying the numerical value of each rating by its relative frequency of use and summing over ratings. Means of the individual means were 3.74 and 3.05 for HP and LP groups, respectively. The individual means were rank-ordered and a Mann-Whitney *U* test showed the difference between groups to be significant ($p <$ .02). From this it may be concluded that, on the average, LP *S*s tended to use lower confidence ratings than did HP *S*s. This may also be seen from the distributions of ratings for the two groups shown in Fig. 1.

It may be noted that, although the LP distribution clearly is shifted toward the low end of the scale relative to the HP distribution, LP *S*s, as a group, used

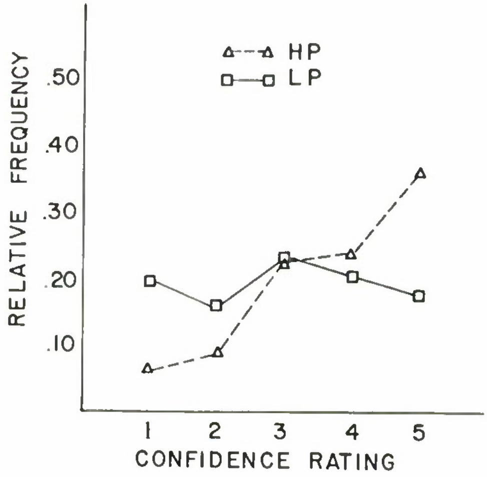

FIG. 1. Distribution of confidence ratings for high and low performance groups

the two highest confidence ratings at least as frequently as the two lowest in spite of their near chance performance on the primary task. It must be emphasized, however, that the nearly rectangular distribution shown in Fig. 1 was not characteristic of individuals within the group. In fact for every *S* of both groups the departure from a rectangular distribution of ratings was significant. [For all *S*s but 1, $\chi^2 > 14$ ($p < .01$); for the remaining *S*, $\chi^2 > 12$ ($p < .02$).] Twenty of the 24 *S*s of each group used one end of the confidence scale (two highest or two lowest ratings) at least 3 times as frequently as the opposite end. Whereas most (21) of the HP *S*s used the high end of the scale, the LP group was fairly evenly divided between *S*s who tended to use one end of the scale and those who tended to use the other; hence the near rectangular distribution of Fig. 1.

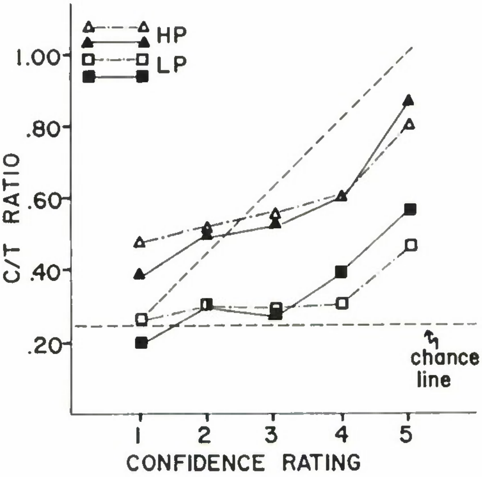

FIG. 2. Ratio of number correct to total number of judgments associated with each confidence rating for each performance group

Ratios of correct-to-total (C/T) number of judgments associated with each confidence rating were computed and are displayed in Fig. 2. The solid lines represent C/T ratios computed from pooled data; the broken lines represent averages of the C/T ratios of individual *S*s. The two procedures appear to give fairly similar plots in this case; however, they could give very dissimilar ones. With pooled data, an individual *S*'s influence in the determination of a particular C/T ratio depends on how frequently (relative to other *S*s) he made use of the associated rating. In the alternative case each *S* is an equally important determinant of each C/T ratio irrespective of the frequency with which he used the different ratings (except in cases in which he failed to use a particular rating at all). This procedure has the unfortunate property that a C/T ratio of 1.00 based on, say, 50 responses of one *S* may be offset by a C/T ratio of 0 based on a single response of another *S*.

As suggested by Fig. 2 and borne out by a test of trend (Hayes, 1957), C/T ratio tended to increase with confidence level for both groups ($p < .01$). However, in neither group were differences in confidence accompanied by proportional differences in performance, except possibly at the upper end of the confidence scale.

If the lowest and highest ratings are interpreted as predictions of chance and perfect performance, respectively (which is consistent with *S*'s instructions), and the rating scale is treated as an equal interval scale, then the diagonal line represents the ideal relationship between C/T ratio and expressed confidence. That is, points lying on the diagonal represent maximal agreement between performance "predicted" from the confidence assignments and that in fact obtained. (Actually perfect agreement is not to be expected even given an ideal assignment of the confidence ratings since they are restricted to five discrete values whereas C/T ratio is not; however, the error due to this fact would be sufficiently slight to be inconsequential in the present context.) Points below the diagonal are suggestive of overconfidence in the sense that the associated confidence ratings predict better performance than was actually obtained. Conversely, points above the diagonal suggest underconfidence since the obtained performance was better than that predicted by the ratings.

As measures of realism of confidence expressions Adams and Adams (1961) have developed algebraic and absolute discrepancy scores defined, respectively, as

$$\Sigma(p_i - P_i)\, n_i / \Sigma n_i \qquad \text{and}$$

$$\Sigma\, | p_i - P_i |\, \sqrt{n_i} / \Sigma \sqrt{n_i}\ ,$$

"in which P_i is the percentage correct at confidence p_i and n_i is the number of decisions made with confidence p_i." The algebraic discrepancy score is equivalent to the algebraic difference between mean confidence and the total per cent correct, and gives an indication of general overconfidence or underconfidence. The absolute discrepancy score gives a weighted average absolute difference between per cent correct observed and that predicted by the confidence assignments. Although, in this experiment, confidence ratings were defined in terms of p_i only in the sense that the ends of the scale were explicitly associated with chance and certainty, for purposes of analysis the five ratings were replaced with p_i ranging in equal steps from 25 (chance) to 100. (Whether comparable results would be obtained if *S*s were required to actually express confidence in terms of probabilities or expected percentages is an empirical question and can be determined only by further experimentation.)

Discrepancy scores were computed for each *S* of both groups. All but 5 *S*s, each of whom was in the HP group, had positive algebraic scores suggesting general overconfidence. The mean algebraic discrepancy scores for HP and LP groups, respectively, were 11.6 and 29.8; mean absolute discrepancy scores for the two groups were 22.4 and 33.5. Mann-Whitney *U* tests showed the between groups differences to be significant in both cases ($p < .01$). It appears that, in terms of the measures proposed by Adams and Adams, *S*s who performed best on the primary task tended to assign confidence ratings more realistically than did those who performed poorly.

Some caution is necessary in interpreting these measures since both may vary strictly as a function of performance on the primary task. As an extreme but convincing illustration, consider the case of two hypothetical *S*s, one of whom is considerably better than the other with respect to the primary decision task, but both of whom assign confidence judgments by pulling them out of a hat. Such an experiment might yield relationships between per cent correct and confidence similar to those illustrated in Fig. 3. Line *a* represents better performance on the primary decision task than does line *b*, but both represent a random assignment of confidence ratings. However, the data represented by line *a* would yield a considerably smaller algebraic or absolute discrepancy score, than would those represented by line *b*. It might be argued that even in this case it is not unreasonable to speak of degrees of realism in the confidence assignments since indeed the correspondence between obtained performance and that predicted by the ratings, however the latter were arrived at, is unquestionably greater in one case than in the other. However, this would seem to stretch the

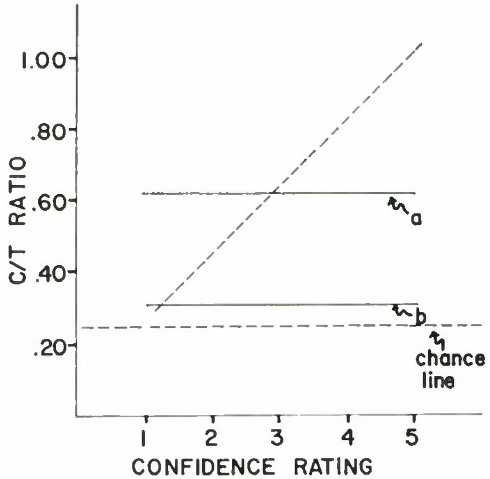

FIG. 3. Plot of hypothetical data illustrating the possibility of differences in discrepancy scores resulting solely from differences in level of performance on the primary task

connotation of realism somewhat beyond its accepted domain and moreover does not seem to be consistent with Adams and Adams' use of the word.

Perhaps a distinction should be made between realism, as defined by Adams and Adams, and what might be called sensitivity; the former denoting the degree of correspondence between predicted and obtained performance measures in terms of absolute values, and the latter reflecting simply the extent to which differences in confidence are indicative of differences in performance on the primary task. With reference to the data represented in Fig. 2, sensitivity is reflected roughly in the slope of a curve; whereas, realism, as measured by Adams and Adams, varies with both slope and intercept. As a crude test for between-groups differences in sensitivity, differences in C/T ratio associated with step increases in confidence were obtained for each *S*. For each step increase in confidence the C/T differences were rank-ordered and a Mann-Whitney *U* test was made on the ranks. Significant between-groups differences were obtained only in the case of performance increments associated with the confidence increment 4 to 5 ($p < .05$). It must be remembered, however, that the probability of getting $p < .05$ by chance in at least one out of 4 different tests is considerably greater than .05; hence, the significance of the difference obtained is questionable. Although we are not justified on the basis of a Mann-Whitney test to conclude that differences do not exist even in those cases in which $p <$ .05 was not obtained, inspection of Fig. 2 suggests that whatever between-group differences there may be are not very large. The plots look quite similar except for differences in intercept.

REFERENCES

Adams, J. K., & Adams, P. A. Realism in confidence judgments. *Psychol. Rev.*, 1961, 68, 33-45.

Carterette, E. C., & Cole, M. *A comparison of the receiver operating characteristics for messages received by ear and by eye.* Dept. of Navy, ONR, Report No. 2, June, 1959.

Hayes, J. R. M. A non-parametric test of trend. *Psychol. Newsltr*, 1957, 9, 29-34.

Henman, V. A. C. The relation of the time of a judgment to its accuracy. *Psychol. Rev.*, 1911, 18, 186-201.

Nickerson, R. S., & McGoldrick, C. C. Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Percept. mot. Skills*, 1963, 17, 159-167.

Pollack, I., & Decker, L. R. Confidence ratings, message reception and the receiver operating characteristics. *J. Acoust. Soc. Amer.*, 1958, 30, 286-292.

Accepted December 10, 1964.